Data Science Capstone Project – Gopal Krishnan

Objective: identify suitable locations for a South Indian restaurant in Hartford, Ct.

Problem Background

I've lived in Hartford the last 20years. In recent years, I've felt the desire for a career change, specifically starting my own restaurant, catering South Indian food. I've noticed that there is not very many South Indian food catering options in the area, while the South Indian demographic has steadily grown in the last few years. Since Asian and indian grocery locations are necessary venues for Indians, locating a South Indian restaurant very near such grocery stores is the chosen approach.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new South India restaurant in Hartford, Connecticut. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: where should an entrepreneur consider opening a South Indian restaurant in Hartford, Ct

Target Audience

Anyone interested in opening a South Indian restaurant in Hartford, Ct.

Data

To solve this problem, I will need the below data:

- List of towns in Hartford county, Connecticut.
- Latitude and Longitude of these towns.
- Venue data related to Indian and Asian grocery stores and restaurants. This will help us find the towns that are most suitable to open the restaurant.

Extracting the data

- Scrapping of Hartford neighborhoods via https://www.geonames.org/postal-codes/US/CT/003/hartford.html to get latitudes and longitudes
- https://en.wikipedia.org/wiki/Hartford,_Connecticut to get population and per capita income information.
- Using Foursquare API to get venue data related to these neighborhoods

Methodology

Data from the above sites for the latitudes and longitudes were first scraped and cleaned and then uploaded as a .csv file to my Github location, from where they were referenced in my Notebook. The outcome looks like this:

	Town	Latitude	Longitude
0	Avon	41.80	-72.83
1	Berlin	41.62	-72.75
2	Bloomfield	41.83	-72.74
3	Bristol	41.68	-72.94
4	Burlington	41.77	-72.96
5	Canton	41.83	-72.90

Next, the demographics data on population, population density, and per capita income were similarly cleaned:

	Town	Per capita income	Population	Pop. Density
0	Avon	66862	22290	781
1	Berlin	38134	19866	736
2	Bloomfield	39738	20486	779
3	Bristol	29629	60477	2257
4	Burlington	43392	9301	306
5	Canton	46401	10292	412

These 2 dataframes were then joined.

	Town	Latitude	Longitude	Per capita income	Population	Pop. Density
0	Avon	41.80	-72.83	66862	22290	781
1	Berlin	41.62	-72.75	38134	19866	736
2	Bloomfield	41.83	-72.74	39738	20486	779
3	Bristol	41.68	-72.94	29629	60477	2257
4	Burlington	41.77	-72.96	43392	9301	306

Next Foursquares was used to start exploring these towns in Hartford County. Sample results for the town of Avon are shown below:

	Town	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Avon	41.80	-72.83	Elephant Trail	41.808008	-72.828137	Thai Restaurant
1	Avon	41.80	-72.83	Max a Mia	41.807845	-72.824397	Italian Restaurant
2	Avon	41.80	-72.83	Bruegger's Bagel Bakery	41.807357	-72.827398	Bagel Shop
3	Avon	41.80	-72.83	The UPS Store	41.807955	-72.827703	Shipping Store
4	Avon	41.80	-72.83	SUBWAY	41.808727	-72.828960	Sandwich Place
5	Avon	41.80	-72.83	Dunkin'	41.807046	-72.824549	Donut Shop
6	Avon	41.80	-72.83	Pick and Mix	41.807818	-72.827272	Korean Restaurant
7	Avon	41.80	-72.83	Countryside	41.801062	-72.824035	Trail
8	Avon	41.80	-72.83	My Dog's Daycare/Doggy Do's	41.797072	-72.836619	Pet Store
9	Avon	41.80	-72.83	Avon Hair Company	41.799881	-72.819965	Health & Beauty Service
10	Avon	41.80	-72.83	Avon House Painting by Franklin	41.802233	-72.839800	Construction & Landscaping
11	Avon	41.80	-72.83	Farmington River	41.806096	-72.823735	River
12	Avon	41.80	-72.83	Carmen Anthony Fishhouse	41.807446	-72.826834	Seafood Restaurant
13	Avon	41.80	-72.83	Avon Cider Mill	41.801753	-72.819343	Farmers Market
14	Avon	41.80	-72.83	Welcome Wine & Liquor	41.807975	-72.827813	Wine Shop
15	Avon	41.80	-72.83	Hot Heaven Pizza	41.807975	-72.827813	Pizza Place
16	Avon	41.80	-72.83	Cake Gypsy	41.808084	-72.827891	Bakery
17	Avon	41.80	-72.83	Little Silver Shop	41.808514	-72.828764	Jewelry Store
18	Avon	41.80	-72.83	Village Garage and Tire Center	41.808786	-72.829828	Auto Workshop

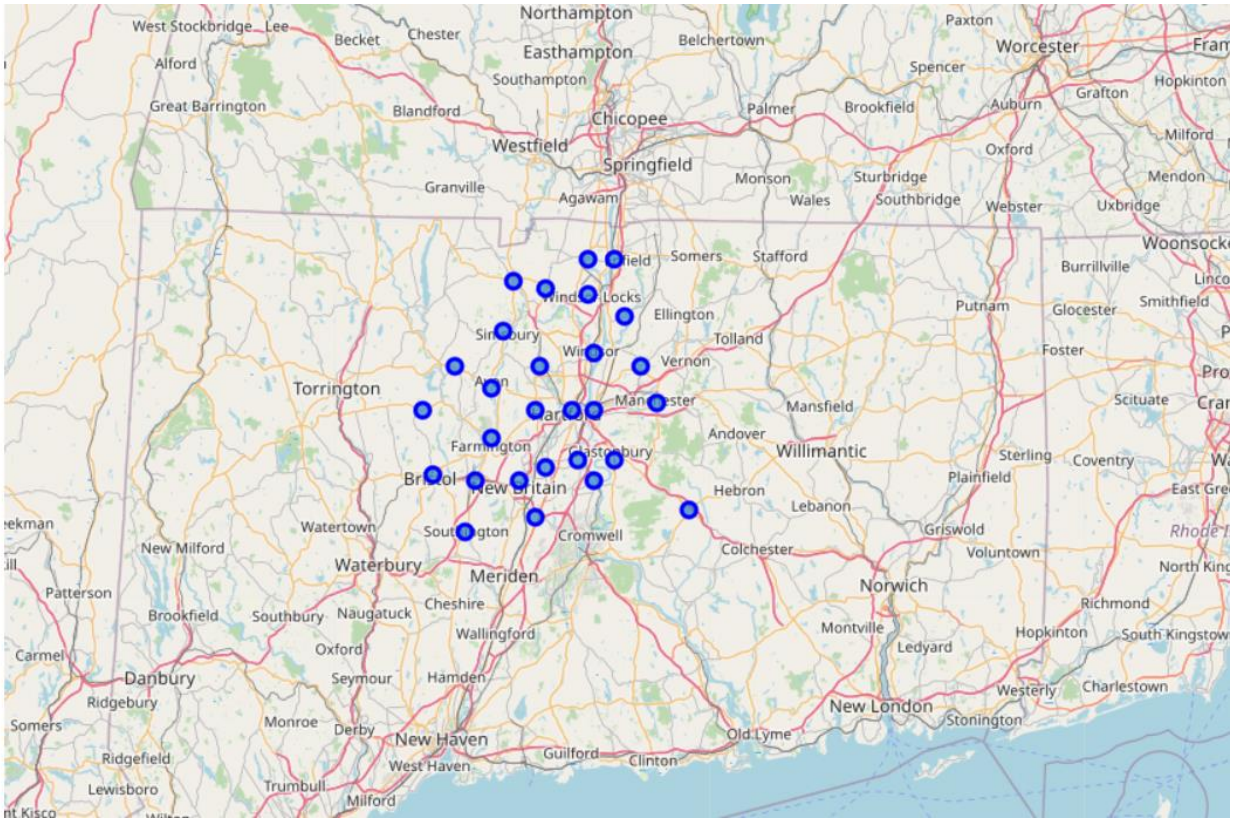
From here, the next step is to identify the number of Indian restaurants in each of these towns. The result is as follows:

Town	Number of Indian restaurants
Bloomfield	2
Bristol	1
Burlington	1
Canton	1
East Hartford	1
Farmington	1
Granby	3
Hartford	2
Marlborough	1
New Britain	3
Plainville	1
Rocky Hill	1
Suffield	1
Windsor	2

This can now be merged with the original dataframe to get:

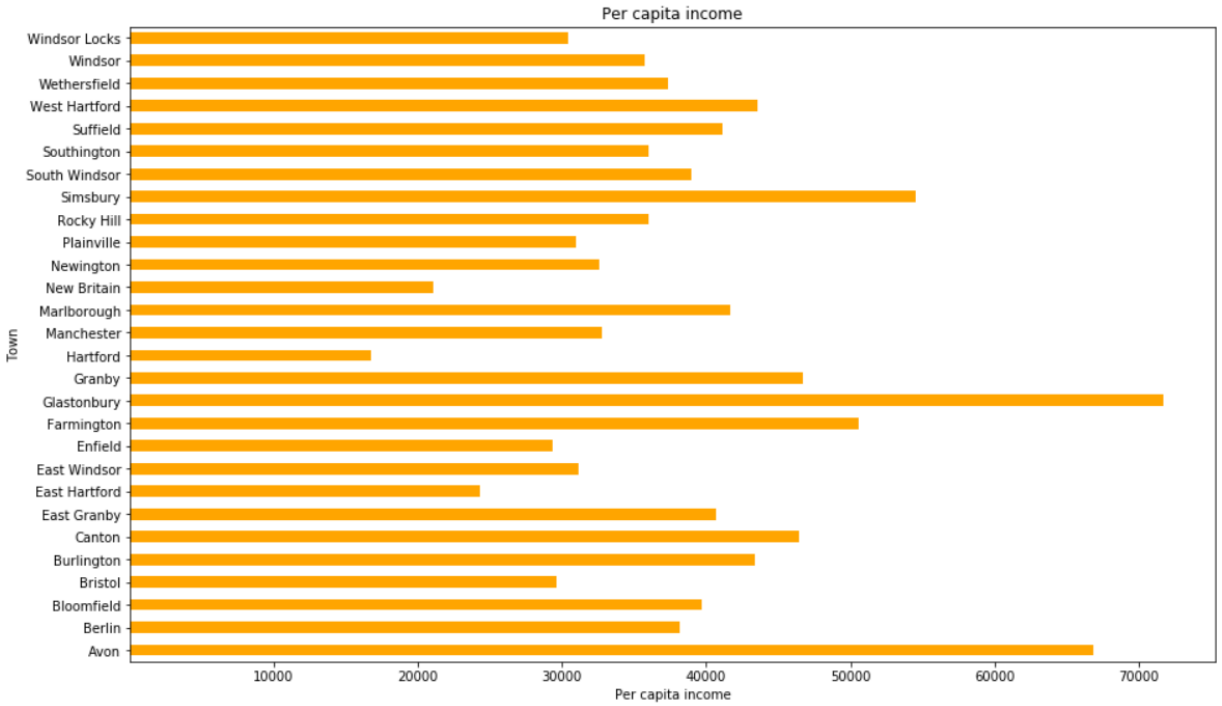
	Town	Latitude	Longitude	Per capita income	Population	Pop. Density	Number of restaurants	Number of Indian restaurants
0	Avon	41.80	-72.83	66862	22290	781	4.0	0.0
1	Berlin	41.62	-72.75	38134	19866	736	0.0	0.0
2	Bloomfield	41.83	-72.74	39738	20486	779	13.0	2.0
3	Bristol	41.68	-72.94	29629	60477	2257	6.0	1.0
4	Burlington	41.77	-72.96	43392	9301	306	2.0	1.0
5	Canton	41.83	-72.90	46401	10292	412	5.0	1.0

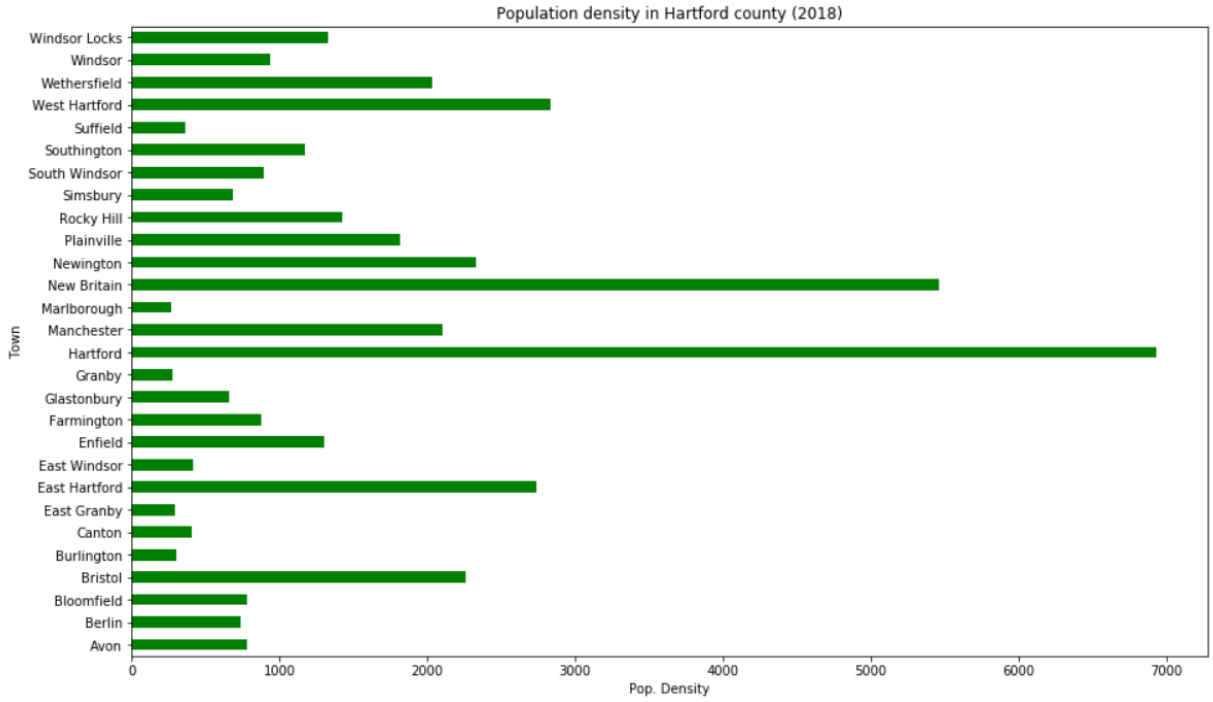
After cleaning and preparing the data, let us identify the steps that have to be performed in order to find the best towns. First, we will apply some basic exploratory analysis to our data. For that let's find the location of each town on the map. Then we can visually inspect some values in our data with the help of bar charts. Secondly, we have the possibility to reduce the number features in data frame by replacing them with more reasonable data. Finally, we will perform cluster analysis to find the best cluster of towns with meaningful features.



Descriptive statistics of the data thus far is shown below.

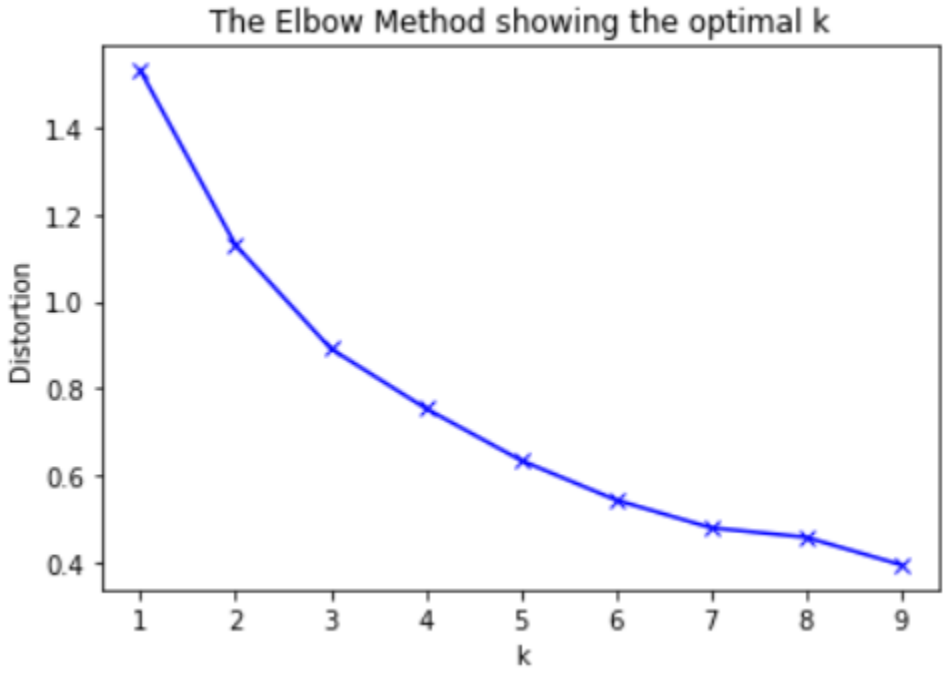
	Latitude	Longitude	Per capita income	Population	Pop. Density	Number of restaurants	Number of Indian restaurants
count	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000	28.000000
mean	41.782857	-72.720000	38866.750000	32003.285714	1516.071429	4.285714	0.750000
std	0.112969	0.128351	12035.028568	25989.281608	1543.399877	3.933306	0.927961
min	41.600000	-72.960000	16798.000000	5148.000000	272.000000	0.000000	0.000000
25%	41.687500	-72.815000	31121.500000	14925.750000	599.000000	1.750000	0.000000
50%	41.770000	-72.730000	37731.500000	24425.500000	916.500000	4.000000	0.500000
75%	41.857500	-72.640000	43427.500000	43465.250000	2052.750000	5.000000	1.000000
max	41.980000	-72.460000	71709.000000	124775.000000	6932.000000	18.000000	3.000000

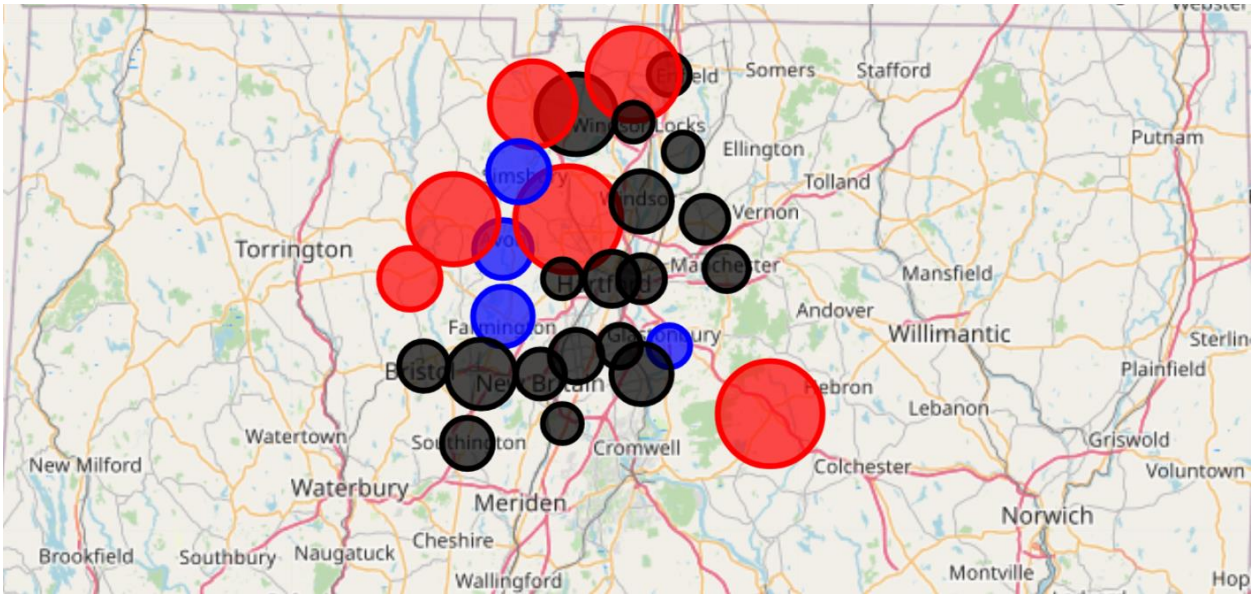




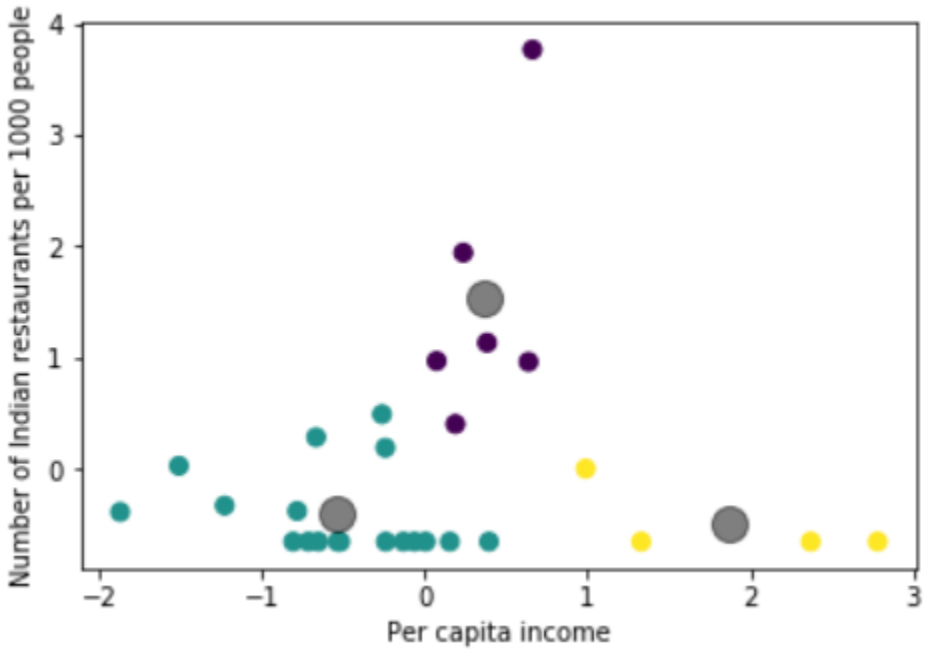
Cluster Analysis

In order to identify groups (clusters) with similar characteristics, let's us apply the unsupervised learning method to our data, namely K-Means algorithm. But before that, we can reduce the number of features and remove columns "Population", "Number of restaurants" and "Number of Indian restaurants". These three columns we can replace with two new ones, namely, "Number of restaurants per thousand people" and "Number of Indian restaurants per thousand people".





Results



	Cluster Labels	Town	Latitude	Longitude	Per capita income	Population	Pop. Density	Number of restaurants	Number of Indian restaurants
2	0	Bloomfield	41.83	-72.74	39738	20486	779	13.0	2.0
4	0	Burlington	41.77	-72.96	43392	9301	306	2.0	1.0
5	0	Canton	41.83	-72.90	46401	10292	412	5.0	1.0
12	0	Granby	41.95	-72.79	46687	11282	277	5.0	3.0
15	0	Marlborough	41.63	-72.46	41669	6404	272	4.0	1.0
23	0	Suffield	41.98	-72.65	41098	15735	366	8.0	1.0

	Cluster Labels	Town	Latitude	Longitude	Per capita income	Population	Pop. Density	Number of restaurants	Number of Indian restaurants
0	2	Avon	41.80	-72.83	66862	22290	781	4.0	0.0
10	2	Farmington	41.73	-72.83	50541	25340	881	5.0	1.0
11	2	Glastonbury	41.70	-72.60	71709	34427	660	1.0	0.0
20	2	Simsbury	41.88	-72.81	54571	23511	685	5.0	0.0

	Cluster Labels	Town	Latitude	Longitude	Per capita income	Population	Pop. Density	Number of restaurants	Number of Indian restaurants
1	1	Berlin	41.62	-72.75	38134	19866	736	0.0	0.0
3	1	Bristol	41.68	-72.94	29629	60477	2257	6.0	1.0
6	1	East Granby	41.94	-72.73	40698	5148	291	2.0	0.0
7	1	East Hartford	41.77	-72.64	24373	51252	2741	4.0	1.0
8	1	East Windsor	41.90	-72.58	31162	11162	416	0.0	0.0
9	1	Enfield	41.98	-72.60	29340	44654	1306	1.0	0.0
13	1	Hartford	41.77	-72.68	16798	124775	6932	18.0	2.0
14	1	Manchester	41.78	-72.52	32752	58241	2103	3.0	0.0
16	1	New Britain	41.67	-72.78	21056	73206	5463	7.0	3.0
17	1	Newington	41.69	-72.73	32561	30562	2333	4.0	0.0
18	1	Plainville	41.67	-72.86	31000	17716	1814	5.0	1.0
19	1	Rocky Hill	41.67	-72.64	36021	19709	1426	4.0	1.0
21	1	South Windsor	41.83	-72.55	38945	25709	896	2.0	0.0
22	1	Southington	41.60	-72.88	36053	43069	1177	5.0	0.0
24	1	West Hartford	41.77	-72.75	43534	63268	2837	0.0	0.0
25	1	Wethersfield	41.70	-72.67	37329	26668	2036	1.0	0.0
26	1	Windsor	41.85	-72.64	35780	29044	937	6.0	2.0
27	1	Windsor Locks	41.93	-72.65	30436	12498	1330	0.0	0.0

Discussion

During the analysis, three clusters were defined. No clear outliers were seen. Two other groups were clustered according to the per capita income. It is obvious that the cluster with highest average income per person could have the highest priority for us (Cluster 2).

Avon, Glastonbury and Simsbury are all very attractive options in terms of distances to the center of their own cluster and relatively high value of income per person (~$60k). They are also very similar in terms of population density (~800), so any of these 3 towns would work for a Indian / South Indian restaurant.

A second way to look for a location is to target very high density areas with not so significantly lower per capita income. Reviewing cluster 1, we see that one excellent location is West Hartford, with a per capita income of $43k (~25% lower than cluster 1 average) but a population density of 2837 (> 200% more than cluster 1 average).

I'd not set up the restaurant in cluster 0 towns due to their low per capita income and low population density.

Reviewing this analysis, West Hartford would be my #1 choice for locating the Indian/South Indian restaurant.

In terms of what could be done further to improve this analysis, it'd be interesting to compare the results of venues from Foursquare to another map, such as Google map or Openstreet map. Maybe more importantly, I've also ignored demographic mixes in this analysis.

Conclusion

To conclude, the basic data analysis was performed to identify the most optimal towns for the placement of the Indian/South Indian restaurant in Hartford county. During the analysis, several important statistical features of the towns were explored and visualized. Furthermore, clustering helped to highlight the group of optimal areas. Finally, West Hartford won over high income (but low population density) towns of Avon, Glastonbury and Simsbury as the chosen location for greater analysis.